1. The least squares estimator of the slope coefficient in a simple linear regression model is

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

and the sample correlation coefficient has the formula

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Starting with the above expression for the slope coefficient, show that it can be expressed as

$$\hat{\beta}_1 = r_{XY}\frac{s_y}{s_x}$$

where s_y is the standard deviation of Y and s_x is the standard deviation of X.

Since $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$, $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$, then $\frac{s_y}{s_x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

So $r_{xy}\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \hat{\beta}_1$$

2. Data were collected from 120 young adult patients. Let x be age (in years) and Y be satisfaction with health care provider (scale of 0-4, higher score indicates more satisfaction). Suppose you have the following summary statistics:

$\bar{x} = 24.7$
$\bar{Y} = 3.07$
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2379.93$
$\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = 92.41$
$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 49.405$
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 45.82$

a. Obtain the sample correlation coefficient, r_{XY}.
b. Obtain the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i independent $N(0, \sigma^2)$.
c. Provide an interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$, including their units.
d. Suppose that you were working with the centered-predictor model, which we will write as $Y_i = \alpha_0 + \alpha_1(x_i - \bar{x}) + \varepsilon_i$, ε_i independent $N(0, \sigma^2)$. What are the least squares estimators of α_0 and α_1?
e. Provide an interpretation of $\hat{\alpha}_0$ and $\hat{\alpha}_1$, including their units.

a. $r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{92.41}{\sqrt{2379.93}\sqrt{49.405}} = 0.269$

b. $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{92.41}{2379.93} = 0.0388$ points/year

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x} = 3.07 - 0.0388 \times 24.7 = 2.112$ points

c. $\hat{\beta}_1$: for each one year increase of young adult patient's age, the expected satisfaction with health care provider increases by 0.0388 points.

$\hat{\beta}_0$: When the age of young adult patient is zero, the expected satisfaction with health care provider is 2.112 points. However, this is only a meaningful interpretation if $x=0$ is reasonable.

d. Let $x_i - \bar{x} = w_i$, then $Y_i = \alpha_0 + \alpha_1 w_i + \varepsilon_i$. So $\hat{\alpha}_1 = \frac{\sum(w_i-\bar{w})(Y_i-\bar{Y})}{\sum(w_i-\bar{w})^2} = \frac{\sum(x_i-\bar{x})(Y_i-\bar{Y})}{\sum(x_i-\bar{x})^2} = \hat{\beta}_1$

$\hat{\alpha}_0 = \bar{Y} - \hat{\alpha}_1\bar{w} = \bar{Y}$. So $\hat{\alpha}_0 = 3.07$ points $= \bar{Y}$, $\hat{\alpha}_1 = 0.0388$ points/year

e. $\hat{\alpha}_0$: When the age of young adult patient is equal to its mean, i.e. 24.7, the estimated mean for satisfaction with health care provider is 3.07 points.

$\hat{\alpha}_1$: for each one year increase of young adult patient's age, the expected satisfaction with health care provider increases by 0.0388 points.

3. Consider the model $Y_i = \beta_0 + \varepsilon_i$. Using the method of least squares, show that the value of β_0 that minimizes the sum of the squared residuals is $\hat{\beta}_0 = \bar{Y}$.

$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0)^2$, let $f(\hat{\beta}_0) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0)^2$

So $f'(\hat{\beta}_0) = 2n\hat{\beta}_0 - 2\sum_{i=1}^{n} Y_i = 0 \Rightarrow \hat{\beta}_0 = \bar{Y}$. $f''(\hat{\beta}_0) = 2n \geq 0$.

When $\hat{\beta}_0 \leq \bar{Y}$, $f'(\hat{\beta}_0) \leq 0$, $f(\hat{\beta}_0)$ is a decreasing function;

When $\hat{\beta}_0 \geq \bar{Y}$, $f'(\hat{\beta}_0) \geq 0$, $f(\hat{\beta}_0)$ is an increasing function.

So $\hat{\beta}_0 = \bar{Y}$ minimizes the sum of squared residuals.

Fit each of the following simple linear regression models and provide an interpretation of the regression coefficient associated with the x variable.

a. Y = risk, x = beds
b. Y = risk, x = svcs
c. Y = nurses, x = age
d. For medical school affiliation, create a variable called med that equals 1 for yes and 0 for no and fit the model with Y = nurse and x = med.

a. $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 3.72404 + 0.0025 x_i$

So an interpretation of the regression coefficient associated with the x variable is: For every one increase in the average number of beds in hospital during study period, the infection risk, i.e. the average estimated probability of acquiring infection in hospital * 100, will increase by 0.0025 on average.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.72404	0.19517	19.08	<.0001
beds	1	0.00250	0.00061579	4.06	<.0001

b. $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 2.78402 + 0.0364 x_i$

So an interpretation of the regression coefficient associated with the x variable is: For every one increase in the available facilities and services, i.e. percent of 35 potential facilities and services that are provided by the hospital, the expected infection risk will increase by 0.0364.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.78402	0.34882	7.98	<.0001
svcs	1	0.03640	0.00763	4.77	<.0001

c. $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 311.06760 - 2.58905 x_i$

So an interpretation of the regression coefficient associated with the x variable is: For every one increase in average age of patients (in years), the expected average number of full-time equivalent nurses during study period will decrease by 2.58905.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	311.06760	157.71376	1.97	0.0511
age	1	-2.58905	2.95251	-0.88	0.3824

d. $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 138.92708 + 228.13174 x_i$

So an interpretation of the regression coefficient associated with the x variable is: If the hospital has a medical school affiliation, the expected average number of full-time equivalent nurses during study period will be 228.13174 more than the hospitals without a medical school affiliation.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	138.92708	11.54610	12.03	<.0001
med	1	228.13174	29.76803	7.66	<.0001